The Ohio State University

The Department of Electrical and Computer Engineering

A thesis presented for the degree of Undergraduate in Electrical and Computer Engineering

Identifying Child Isolation in Preschool Classrooms using Computer Vision Techniques

Author
Joseph Chiu

Supervisor
Dr. Kevin Passino

April 22, 2020

Abstract

The advent of modern sensing technologies has allowed data collection on people and events to occur at an unprecedented scale. One such application has been to study different facets of early childhood behavior. Research has shown that when children first start compulsory education, their environment and interactions with peers and figures of authority can either boost their development, or become a basis for negative outcomes. Specifically, peer-to-peer isolation can occur as early as 3-5 years of age, and can be especially devastating to developing children. This thesis outlines the process of designing an experiment to collect video data from a preschool classroom, and develop scientific, quantitative methods for studying isolation, using computer vision techniques. While this phenomenon is easy to understand, it is challenging to rigorously define and identify. Preliminary results here show that an "isolation score" that is based on computer vision data is a promising method for identifying child isolation in a classroom.

Contents

1 Introduction

Nearly 80% of children between the ages of 3-5 spend a majority of their time in out-of-care settings such as preschools. These children are experiencing peak development of numerous brain systems, including ones for language and cognition skills, and the ability to socialize and interact with peers, and studies have shown that the environment greatly impacts mental development during these formative years.

One of the main pillars of childhood development is social interaction, where children face the challenge of adapting to out-of-home environments, while concurrently learning to form relationships with peers and teachers. Starting compulsory education can introduce especially stressful situations for children trying to adapt to new social environments. Just as positive influences such as social acceptance from peers, and warm interactions with figures of authority can be greatly beneficial to a child's emotional and cognitive development, the reverse, social isolation, can be equally damaging, and even leave persistent effects [1]. It is even suggested that peer rejection may result in diminished respect for authority and prosocial behavior, among other negative outcomes [2].

However, traditional methods of studying human social interactions rely heavily on trained human observers to hand-code specific actions and behavior, limiting the amount of data that can be obtained, and the complexity of patterns that can be studied. The possibility of children reacting to an unusual adult observer presence also adds variance to observations. To this end, researchers have recently begun employing modern sensing technologies to automate collection of data from classrooms [3, **?**, 4] . While the goals of these research groups are different, from studying movement during child development, to quantifying engagement in a college classroom, the common thread involves using non-invasive, automated, techniques to collect data about the dynamics of a classroom, and developing computational methods to extract meaningful information.

Isolation was selected as a phenomenon to further investigate due to the expertise and research interests of one of our primary collaborators at the Crane Center for Early Childhood Development and Policy, Dr. Laura Justice. To this end, a longitudinal approach to automate the collection of real time location information for individual children in kindergarten classrooms was devised. Unlike most research groups studying similar problems, we opted to use cameras to provide localization information, rather than methods based on radio-frequency identification (RFID) technology. The ability to provide greater localization accuracy, compared to RFID, coupled with the richer data that video encodes were the main drivers for this decision.

In reality, child isolation is a complex concept that has different definitions and effects on individuals. For example, early childhood researchers have identified multiple classes of peer-to-peer interactions, some of which fall under the umbrella of isolation [4]. However, these interactions were all completely hand-coded by human researchers spending hours to watch camera footage, which is not feasible for large-scale experiments or datasets. This paper outlines the concept and implementation of an "isolation score", a method to identify chil-

dren that are physically isolated using computer vision techniques. Applications for this isolation score include providing instructors with real-time updates for the children in their classroom, posteriori analysis on the causes and motion of children, and even the monitoring of clusters of children to prevent the spread of COVID-19 in classrooms.

2 Methods

2.1 Setting and Participants

As data collection involved video recording human subjects, an Ohio State University IRB was applied for and approved. The data collection portion of the experiment took place at Ohio State's Schoenbaum Family Center (SFC), in a preschool classroom. Consent was solicited from the guardians of all 20 children present in the room during data collection, along with 3 full-time instructors. The demographics of the children participating reflects the wide diversity of students at SFC, with 9 boys and 10 girls. 60% were African-American, and 40% were White.

2.2 Procedures

Over the course of one week, from 2/11/2020–2/17/2020, two hours of video were recorded each day, starting at 11:00am and 3:00pm, respectively. This ensured the capture of different activities throughout the school day.

In order to automate the process of producing quantitative studies of classroom movement dynamics in a classroom, a pair of GoPro cameras was used to capture video during pre-determined times. The position of the cameras was chosen to provide the greatest field-of-view of the classroom, easy manual access, and robustness. The choice of camera and position was strongly influenced by the characteristics of the specific classroom at SFC. Lack of external power sources higher up on the walls made installation of wired cameras difficult, and limited possible mounting locations. The cameras also had to be placed in such a way that children could not tamper with the system, either intentionally or unintentionally.

Both GoPro cameras were manually activated by an on button, and had enough SD Card storage to last throughout the entire week of data collection. Battery packs plugged into the cameras provided enough charge to last through an entire day of collection, but needed to be recharged every night.

At the beginning of the data collection period, a 8×7 grid with 5.23 in squares was used to calibrate both cameras.

3 Computer Vision Algorithms

While video provides a richer dataset compared to RFID, one of the main drawbacks is that the detection, localization, and identification of a human in a frame

does not come for free. To this end, a pipeline to generate data with semantic meaning from raw video was developed.

The following analysis was all performed using a 12 minute video sequence from the first day of testing. To decrease computational time, the video was sampled at 250 ms, resulting in 2841 total frames processed, and 404 unique tracks identified.

3.1 Detection

The first step in the pipeline was the detection of humans in each video frame. Object detection is a well-studied problem in computer-vision, and many solutions exist. As both the children and teachers exhibit highly dynamic behavior, moving around from different tables and engaging in different tasks, a robust system that can detect people with different orientations, scales, and occlusions was necessary.

Two popular deep-learning networks for object detection were explored for this purpose: YOLOV3 [5], and Faster-RCNN [6]. While both networks generate the same output, a bounding box for each detection in a frame, the implementations are completely different. YOLOV3 provides extremely fast detect ions, with Faster-RCNN about 50 times slower when operating on the same input. However, Faster-RCNN was much more consistent in the number of detections. With 20 children and 3 adults in the classroom during data-collection, the average number of detections per frame by Faster-RCNN was much higher, compared to YOLOV3. It is theorized that this difference is due to the high number of occlusions occurring during the video, often lasting for 30 seconds or longer, as children across from each other at tables create occlusions. The implementation of Faster-RCNN region proposals seem to make it more suitable to the task of detecting targets close in size and proximity, with a lot of overlap. To improve inference times of the detection network, the input image was cropped, down sampled spatially, and also temporally. With YOLOV3, inference time was not a concern. However, with Faster-RCNN, inference of a 12 min input video would take hours, even when downsampled. Temporal downsampling was done by using every seventh frame for detection, or 250 ms. This value was chosen to improve inference time without sacrificing much temporal-spatial information, the importance of which will be discussed in the following section.

3.2 Unique Identification and Tracking

The second step in the pipeline is the unique identification of detections generated by the neural network. Inputs to the detection stage of the pipeline, Faster-RCNN, were 1920×820, a cropped version of the original frame size to exclude wall/ceiling space where no detections could appear. The outputs were rectangular bounding boxes outlining the estimated position of a person, and do

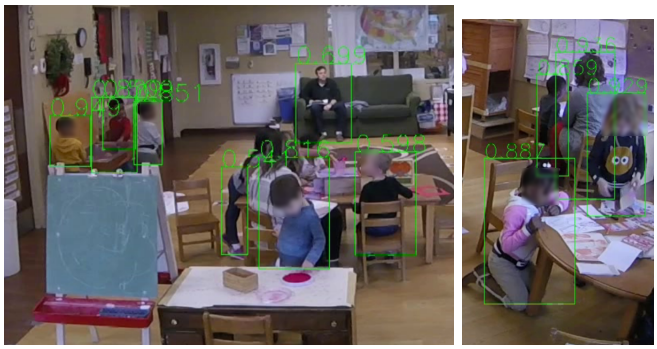
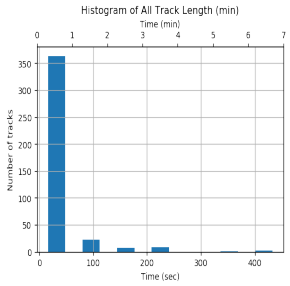


Figure 1: Examples of detection of people using Faster-RCNN, which performs admirably even in highly-occluded and overlapping scenes.

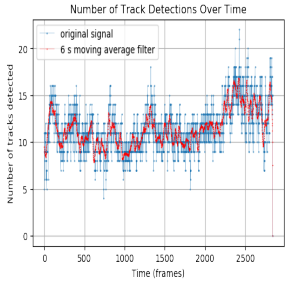
not contain any information differentiating the boxes. Unique identification is an important component in the study of isolation, in order to study the patterns of specific individuals. Early on, different fiducial tags, such as April Tags [7] were explored to facilitate unique identification. Although the extremely low false-positive rate of April Tags was promising, issues in detecting the tags at scales up to the length/width of the classroom ultimately lead the idea to be discarded.

This problem of unique identification of multiple unknown individuals is known as the "Multiple Object Tracking", and is a nascent problem in computer vision [8]. To this end, a method known as "Simple Online and Real time Tracking with a Deep Association Metric" (SORT)[9, 10] is adopted. This approach fuses information from an object detector (Faster-RCNN), with a basic estimation model of the movement of humans from frame to frame, to develop unique "tracks" that have temporal persistence. These tracks are assigned IDs, that can be thought of as identifications for the individuals in the video. The integration of a "Deep Association Metric" generated through a novel metric learning approach [11] greatly improves the ability of the algorithm to overcome visual differences between frames, and occlusion of targets. This metric is learned by feeding RGB images of the bounding boxes, cropped from the original image, through a feature extraction network, resulting in a vector of length 128. These latent representations of the RGB bounding boxes are then used to learn a metric to differentiate the images. The coupling of this metric with a Kalman filter work to improve associate detections from one frame to another [11].

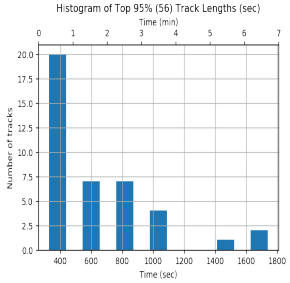
Some characteristics of these tracks are seen in Figures 1 and 2. The addition of temporal persistence to bounding boxes is a step closer to complete unique identification of all participants in a room, but also presents new challenges. Figure 1 shows that most tracks have a very short lifespan, and can be rejected, both on basis of outlier rejection, and dimensionality reduction. By taking the top 10% of tracks, based on length, resulting in 41 nascent tracks, it is more



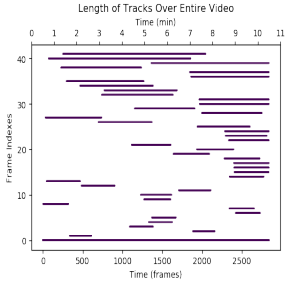
(a) Distribution of all track lengths. Clearly, most tracks do not even last for one minute and can be rejected.



(b) Each frame, the tracker returns a set of detected tracks. One track will not appear more than once in a frame. The red line was calculated using a moving average window of 25 frames, or 6 seconds.



(c) Taking the top 10% of tracks by length results in 41 unique tracks, and on average last x seconds longer compared to (a).



(d) Each horizontal line specifies a single track, from it's creation to termination. The Y axis is simply a categorical index for the different IDs.

Figure 2: Visualizations of the distribution of track lengths across time.

feasible to do both quantitative and qualitative analysis.

4 Determination of Social Isolation and Clustering

4.1 Social Isolation Score

In order to study social isolation, a metric to quantify child isolation was developed. The driving principle behind this metric is that it should be positively

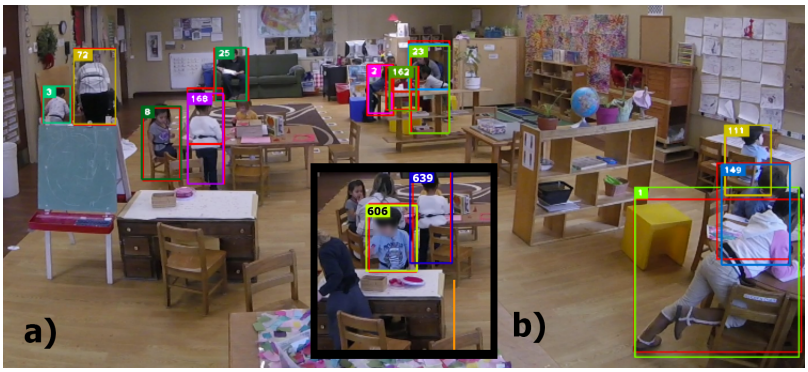


Figure 3: (a) Examples of multiple object tracking using the SORT algorithm. Unlike the Faster-RCNN results in Figure 1, the bounding boxes now have a unique ID (b), which is better shown in the center image. SORT combines a Kalman Filter for updating estimates of track positions, and a feature extractor to better guide track assignment from frame to frame. SORT is able to continue updating a track even if it becomes occluded.

correlated with "isolation," which is loosely defined by a person being physically distanced from others for a certain period of time. Such a metric allows "isolates" to be automatically detected, using data that has passed through the computer vision detection and identification phases mentioned above. The social isolation metric $I(i,t)$ is defined for each track i and time t by

$$I(i,t) = \frac{1}{|P(t)|} \sum_{j \in P(t), j \neq i} F(i,j,t)$$

where the set tracks, minus the i^{th} track, at time t is $P(t)$ and the sum is taken overall all $j \neq i$ that exist at time t. The normalization in front of the sum is used to ensure that $I(i,t) \in [0,1]$, with "0" representing not isolated and "1" representing completely isolated. Also, $F(i,j,t)$ is the degree of non-overlap of tracks i and j at time t,

$$F(i,j,t) = \frac{1}{\sum_{t''=0}^{N} \alpha^{-t''}} \sum_{t'=t}^{t-N} \alpha^{(t-t')} E(i,j,t')$$

where N is the number of steps to consider the degree of exclusion in the past, α is a weight that is sometimes called a "forgetting factor" ($\alpha > 1$) and it weights the importance of being excluded at the current time more than in the past. The factor in front of the sum normalizes so that $F(i,j,t) \in [0,1]$. If $t' = T$ -N_c < 0, then the score is set to zero. This means that for the first N frames of a track, the score will be zero. Also, $E(i,j,t)$ quantifies the exclusion of track i

from track j at time t,

$$E(i,j,t) = \begin{cases} 1 & \text{if } ||c_i(t) - c_j(t)||^2 > \epsilon \\ 0 & \text{else} \end{cases}$$

where $||\cdot||$ is the Euclidean norm, $c_i(t)$ is the centroid of the bounding box for track i at time t, and ϵ is the threshold defining what it means for person i to be isolated from person j at time t.

4.2 Social Clustering Score

The social clustering score $S_c(i,t)$ is defined for each track i and time t by

$$S_c(i,t) = \frac{1}{|P(t)|} \sum_{j \in P(t), j \neq i} F_c(i,j,t)$$

where the set tracks, minus the i^{th} track, at time t is $P(t)$ and the sum is taken overall all $j \neq i$ that exist at time t. The normalization in front of the sum is used to ensure that $S_c(i,t) \in [0,1]$. Also, $F_c(i,j,t)$ is the degree of non-overlap of tracks i and j at time t,

$$F_c(i,j,t) = \frac{1}{\sum_{t''=0}^{N_c} \alpha_c^{-t''}} \sum_{t'=t}^{t-N_c} \alpha_c^{(t-t')} E_c(i,j,t')$$

where N_c is the number of steps to consider degree of exclusion in the past, α_c is a weight, sometimes called a "forgetting factor" ($\alpha_c > 1$), which weighs the importance of being excluded at the current time more than in the past. The factor in front of the sum normalizes so that $F_c(i,j,t) \in [0,1]$. Also, $E_c(i,j,t)$ quantifies the exclusion of track i from track j at time t,

$$E_c(i,j,t) = \begin{cases} 1 & \text{if } ||c_i(t) - c_j(t)||^2 < \epsilon_c \\ 0 & \text{else} \end{cases}$$

where $||\cdot||$ is the Euclidean norm, $c_i(t)$ is the centroid of the bounding box for track i at time t, and ϵ_c is the threshold defining what it means for person i to not have social distancing with j at time t. Here, the parameters can be set different from the isolation case above. This score is higher when are people close together for a longer period of time. Mathematically, the social clustering score is the inverse of the isolation score.

4.3 Tuning Parameters

The isolation metric has three parameters that dictate its interpretation; α, N, and ϵ. These parameters are the forgetting factor, number of past frames to consider, and threshold radius, respectively. As they appear in the same term, the effects of tuning α and N are similar. This relationship can be demonstrated

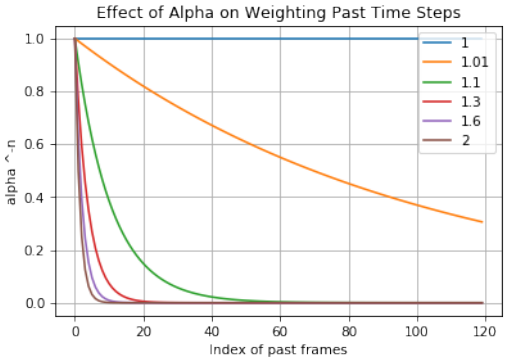


Figure 4: Tuning α essentially also controls the past number of frames considered at a time t. When α is large (> 1.3), the factor $\alpha^{(t-t'')}$ drops within a few frames, and effectively renders any later frames as non-factors.

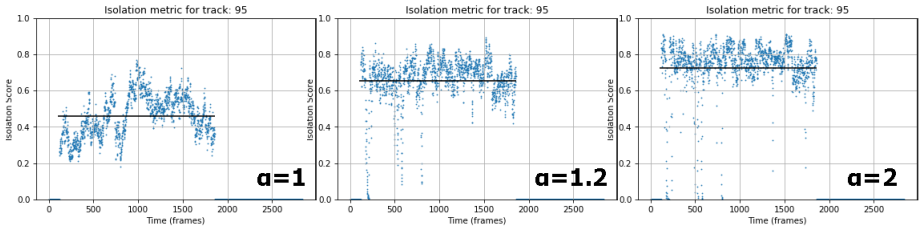


Figure 5: The effects of tuning α, the "forgetting" factor in the isolation score, are observed. The three values tested were 1,1.2, and 2. An α of 1 means that the contribution of each previous frame is the same, while an α of 2 means that the contribution is reduced by a power of 2.

in Figure 4. As α is increased, the contributions of past frames are diminished quickly, rendering a large N ineffective. N can also be adjusted similarly.

Figure 5 shows the effects of α on the isolation metric. A low α (close to 1) coupled with a high N results in a smoother score trajectory with fewer outliers.

The radius ϵ was selected by qualitatively plotting circles around tracks and playing back the video. One major issue with the radius is that it is fixed with pixel values instead of a real world distance. Due to the perspective projection of 3D world points to the image plane, similar to parallel railroad tracks appearing to intersect at the horizon, the fixed radius will include a larger real world distance when people are farther away.

Case	Length (frames)	Mean $I(i,t)$	Median $I(i,t)$	Visual Inspection
Track 37*	705	0.71	0.80	Isolated
Track 95	1784	0.70	0.73	Isolated
Track 590	186	0.35	0.38	Clustered
Track 1282*	661	0.77	0.80	Isolated
Average	647	0.38	0.39	

Table 1: Key statistics for four hand-picked tracks. *Tracks 37 and 1282 correspond to the same child. The average in the bottom row was calculated over the top 10% of tracks, by length.

5 Results

The validity of the isolation score was tested using the unique tracks generated by the SORT algorithm. The time varying isolation score $I(i,t)$ was calculated for each track i.

As this isolation score has not been implemented in such a classroom setting before, there is no "ground truth" or reference data to make comparisons against. The magnitude of the isolation score was qualitatively evaluated by taking tracks with the highest isolation scores, and playing back the original video, with bounding boxes on the target tracks. This was done to ensure the positive correlation between the isolation score and observed "true isolation," as naturally defined by an adult observer. Figures 6-10 show 3 examples of using the isolation metric in conjunction with video to gain insight into the isolation behavior of a specific track.

Figures 6 and 7 correspond to track 37, a child who spent most of his time at a table a lone. While other people passed by, and sometimes sat across from him, no one stayed for more than a few seconds. Although this information can be easily deduced by watching the video, the high isolation score in Figure 7 also suggests such a deduction, but without the need for human interpretation.

Figures 8 and 9 correspond to two different tracks, a rare case where a track of an individual is dropped after a long sustained period, but picked up again. It is a future direction to analyze such cases in more detail.

Figures 10 and 11 represent a track (590) with a very low isolation score, and therefore a higher social clustering score. Video analysis revealed that track 590 corresponds to a teacher sitting at a table surrounded by children in a very socially clustered manner. Again, this example further validates the promise of the social clustering score as a method to *both* automatically identify the physical isolation and social grouping of an individual.

After further individually inspecting tracks and their corresponding video segments, it is clear that high isolation scores sustained over a period of time can be a strong indicator for "true isolation," and possible concern from instructors.

Table 1 Summarizes the key numerical results of the 4 tracks previously mentioned.

Figure 6: Classroom scene, with a bounding box and threshold radius for track 37. As the child is quite distanced from the rest of the children at the table, this would classified as isolation.

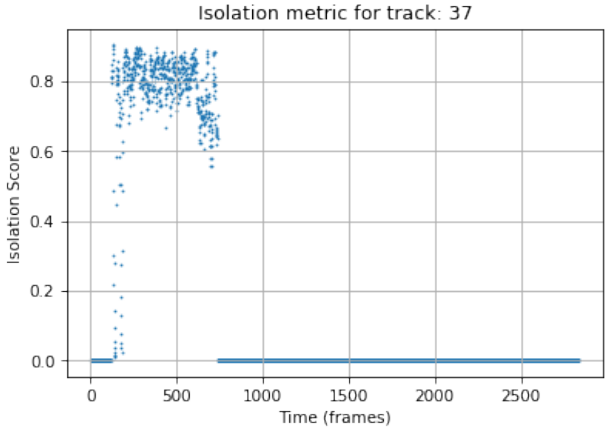


Figure 7: Plot of track 37's isolation score over time.

Figure 8: Classroom scene, with a bounding box and threshold radius for track 1282. As the child is quite distanced from the rest of the children at the table, this would classified as isolation.

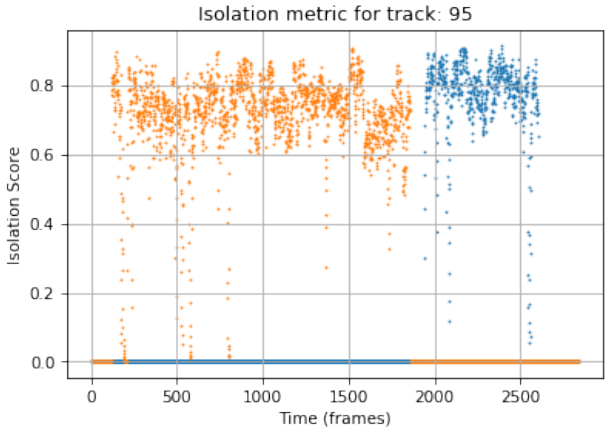


Figure 9: Plot of Tracking 95 and Track 1282's isolation score. An interesting case occurred where the initial track for this child was dropped after a considerable amount of time detected, but then picked up again, with a different ID. The orange points represent the first period of detections, while the blue represents the second period.

Figure 10: Classroom scene, with a bounding box and threshold radius for track 590. Here, the track corresponds to a teacher who is surrounded by students, a case where the isolation score is expected to be low and therefore the social clustering score is high.

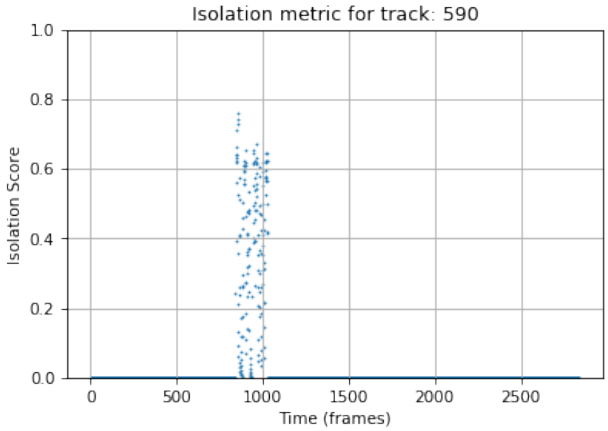


Figure 11: Plot of track 590's isolation score over time showing how social clustering occurs.

6 Conclusion

The isolation score introduced in this paper has shown to be a strong indication of true isolation, a physical and psychological phenomenon that is extremely

difficult to quantify, based on qualitative tests. Playing back the video centered on the five highest scoring tracks (mean of isolation score) clearly identified two individual children who spent long periods of time without extended proximity by either peers or teachers. While there are many directions left to extend the study of this metric with the existing data, the results are promising.

7 Future Work

There are many further extensions for this work. Applying the isolation metric to the remaining video segments will further elucidate the statistical nature, and provide direction on tuning α, N, and ϵ. Establishing a one-to-one mapping from track ID's to study participants is another extremely important portion to the study of isolation, and could significantly boost the utility of the isolation metric. The complement of the isolation score, social clustering, has become an extremely salient idea today, and the data collection and processing methods outlined in this paper could be augmented to study and monitor interactions between children once schools reopen.

8 Acknowledgement

I would like to sincerely thank all the people who supported and guided me throughout my thesis research, and who made this project happen. First and foremost, I wish to acknowledge my appreciation for my advisor Dr. Passino, who mentored me throughout the entire research process, and provided unwavering encouragement. I would like to acknowledge Dr. Justice and the entire research group at the Crane Center and Schoenbaum Family Center, for allowing us use your facility, and for all your practical help and your expertise. I am greatly indebted to Hugo for working tirelessly to get the IRB approved and obtain consent from the parents, as well as setting up the equipment every day during the experiment. Without you, this experiment would not have happened. I would also like to thank Dr. Harry Chao for graciously accepting to be on my defense committee. Lastly, I would like to thank Andrew Fu, my fellow labmate, for his friendship and support over this past year.

References

[1] L. J. Martín-Antón, M. I. Monjas, F. J. García Bacete, and I. Jiménez-Lagares, "Problematic social situations for peer-rejected students in the first year of elementary school," *Frontiers in psychology*, vol. 7, p. 1925, 2016.

[2] L. M. Justice, H. Jiang, and K. Strasser, "Linguistic environment of preschool classrooms: What dimensions support children's language growth?," *Early Childhood Research Quarterly*, vol. 42, pp. 79–92, 2018.

[3] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.

[4] L. J. Chaparro-Moreno, L. M. Justice, J. A. Logan, K. M. Purtell, and T.-J. Lin, "The preschool classroom linguistic environment: Children's first-person experiences," *PloS one*, vol. 14, no. 8, 2019.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[7] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4193–4198, IEEE, 2016.

[8] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim, "Multiple object tracking: A literature review," *arXiv preprint arXiv:1409.7618*, 2014.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649, IEEE, 2017.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, IEEE, 2016.

[11] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 748–756, IEEE, 2018.